
ROBERT TAYLOR

Rein in the four horsemen of irreproducibility

Dorothy Bishop *describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.

In 1975, psychologist Anthony Greenwald noted that science is prejudiced against null hypotheses; we even refer to sound work supporting such conclusions as 'failed experiments'. This prejudice leads to publication bias: researchers are less likely to write up studies that show no effect, and journal editors are less likely to accept them. Consequently, no one can learn from them, and researchers waste time and resources on repeating experiments, redundantly.

That has begun to change for two reasons. First, clinicians have realized that publication bias harms patients. If there are 20 studies of a drug and only one shows a benefit, but that is the one that is published, we get a distorted view of drug efficacy. Second, the growing use of meta-analyses, which combine results across studies, has started to make clear that the tendency not to publish negative results gives misleading impressions.

Low statistical power followed a similar trajectory. My undergraduate statistics courses had nothing to say on statistical power, and few of us realized we should take it seriously. Simply, if a study has a small sample size, and the effect of an experimental manipulation is small, then odds are you won't detect the effect — even if one is there.

It is wasteful to conduct studies that are underpowered, but researchers have often treated statisticians who point this out as killjoys. In 1977, Jacob Cohen wrote a definitive book on the subject; ten years later, another statistician wrote, "Small studies continue to be carried out with little more than a blind hope of showing the desired effect" (R. G. Newcombe *Br. Med. J. (Clin. Res. Ed.)* **295**, 656–659; 1987). In fields such as clinical trials and genetics, funders have forced improvements to working practices by insisting that studies be adequately powered. Other disciplines have yet to catch up.

I stumbled on the issue of *P*-hacking before the term existed. In the 1980s, I reviewed the literature on brain lateralization (how sides of the brain take on different functions) and developmental disorders, and I noticed that, although many studies described links between handedness and dyslexia, the definition of 'atypical handedness' changed from study to study — even within the same research group. I published a sarcastic note, including a simulation to show how easy it was to find an effect if you explored the data after collecting results (D. V. M. Bishop *J. Clin. Exp. Neuropsychol.* **12**, 812–816; 1990). I subsequently noticed similar phenomena in other fields: researchers try out many analyses but report only the ones that are 'statistically significant'.

This practice, now known as *P*-hacking, was once endemic to most branches of science that rely on *P* values to test significance of results, yet few people realized how seriously it could distort findings. That started to change in 2011, with an elegant, comic paper in which the authors crafted analyses to prove that listening to the Beatles could make undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). "Undisclosed flexibility," they wrote, "allows presenting anything as significant."

> MANY RESEARCHERS PERSIST IN WORKING IN A WAY ALMOST **GUARANTEED** NOT TO DELIVER **MEANINGFUL RESULTS.**

The term HARKing was coined in 1998 (N. L. Kerr *Pers. Soc. Psychol. Rev.* **2**, 196–217; 1998). Like *P*-hacking, it is so widespread that researchers assume it is good practice. They look at the data, pluck out a finding that looks exciting and write a paper to tell a story around this result. Of course, researchers should be free to explore their data for unexpected findings — but *P* values are meaningless when taken out of context of all the analyses performed to get them.

The problems are older than most junior faculty members, but new forces are reining in these four horsemen. First, the field of metascience is blossoming, and with it, documentation and awareness of the issues. We can no longer dismiss concerns as purely theoretical. Second, social media enables criticisms to be raised and explored soon after publication. Third, more journals are adopting the 'registered report' format, in which editors evaluate the experimental question and study design before results are collected — a strategy that thwarts publication bias, *P*-hacking and HARKing. Finally, and most importantly, those who fund research have become more concerned, and more strict. They have introduced requirements that data and scripts be made open and methods be described fully.

I anticipate that these forces will soon gain the upper hand, and the four horsemen might finally be slain. ■

Dorothy Bishop *is an experimental psychologist at the University of Oxford, UK.*
e-mail: dorothy.bishop@psy.ox.ac.uk